Personal Statement - Caleb Ziems

Introduction. I have never been afraid to cross disciplinary boundaries. Instead, I actively seek roles that bridge what sociologist Ronald Burt calls "structural holes," straddling different communities and schools of thought so that I might gain new and complementary ideas as well as opportunities to positively impact disparate populations. These goals have guided my academic ambitions, and they are my primary motivation for pursuing a research career: I want to place myself where opportunities for **innovation**, **impact**, and **service** are abundant.

In my work, I develop both insights and algorithms to capture the meaning of natural language *within its social context*. I consider human goals, intentions, relationships, behaviors, and the outcomes of those behaviors for online social groups. This present ambition first took root when I was straddling questions of *language, mind, society, and computation* during my first year of undergrad. It was at a seminar hosted by Center for Mind, Brain, and Culture where Phillip Wolff, professor of psychology, first revealed to me that language can predict human behavior. From tweets alone, he predicted users' long-term health, spending habits, and propensity for risk-taking. Where previous studies were limited by the scale, sample size, and observation window afforded by a laboratory setting, the unique affordances of the web had made it possible to measure these correlations across millions of users. Then and there, I decided I would make the web my laboratory. I would leverage the rapid advances in statistical natural language processing, machine learning, and data mining to answer long-standing questions about human behavior, language, and communication. That week, I emailed Prof. Wolff with a research proposal and set out to join the emerging field of *Computational Social Science*.

Research experience. My intuition was that *language and society inform one another*. In my project with Dr. Wolff, I hypothesized that individuals who are more central in their social networks will also have greater access to new information and will be more likely to discuss upcoming events. To test this hypothesis, I scraped Twitter users' social graphs and counted all future references in their tweets, which I identified computationally using a set of lexical and syntactic tree-structure (Tregex) rules. I found that Twitter users with higher betweenness centralities also tend to share a proportionally greater number of future-oriented tweets, and I presented at Emory University's Undergraduate Research Symposium. Throughout the course of my project, I regularly met with a diverse group of peers to receive critical feedback on both my methods and my ability to present my work. In this way, I developed myself, not only as an **independent researcher**, but also as a **communicator of research ideas**.

By the summer of 2018, my experiences qualified me for an NSF Research Experience for Undergraduates (REU) at Stanford University. I was selected from a pool of over 500 applicants to work alongside 13 other highly motivated interns with backgrounds in computer science, mathematics, psychology, linguistics, communication, and network science. Through stimulating discussions with this group of peers, I discovered a new outlet for my research ambitions. I would use computational methods to *mitigate antisocial behaviors*. That summer, I worked with Dr. Jure Leskovec and Dr. Srijan Kumar to characterize and predict "pump and dump" manipulations in cryptocurrency markets. To do so, I scraped, cleaned and aggregated over 60,000 crypto-related messages, and I acquired data for 177 million trades through a partnership that I helped arrange with a local startup. From this wealth of data, I statistically confirmed the adverse effects that coordinated pumps have on markets, and I trained predictive models to anticipate and prevent these attacks. By the end of the program, I had gained critical skills in **large-scale data processing** and **machine learning** over **evolving and socially-situated language data** with significant real-world impacts. Furthermore, I learned to maintain an **industry partnership** and secure special access to domain-relevant private data.

In the summer of 2019, I proposed a new project in this space, working with Dr. Fred Morstatter at the USC Information Sciences Institute's REU site. This time, I wanted to improve cyberbullying detection systems. Existing annotation schemes lacked standardized labeling criteria and largely disregarded social context, focusing instead on text from individual messages. To solve this, I **developed a new framework** for crowdsourcing annotations where I displayed full message threads in their original format. Additionally, drawing upon the social sciences literature, I decomposed the nuanced problem of cyberbullying into five explicit criteria, and trained annotators accordingly. These criteria were aggressive language, repetition,

harmful intent, visibility among peers, and power imbalance. I showed that existing NLP approaches to cyberbullying detection, like dictionaries, *n*-grams, and word embeddings, can be very effective at detecting swears and other explicitly aggressive language, but they cannot reliably detect the more contextual aspects of cyberbullying like harmful intent. I proposed a new set of social and linguistic features to capture this context, which significantly improved my models' ability to distinguish harmful messages from lighthearted jokes, and to infer users' relative positions of power. I wrote our conference paper as first author, and it received **Honorable Mention for Best Paper Award at ICWSM 2020** [1]. After extending this work in my senior thesis, I received **Highest Honors** as well as the **Academic Excellence Award** in Computer Science.

Most recently, in the summer before starting my PhD program, I started two new hate speech projects. In the first project, my team released a new dataset on the dynamics of anti-Asian hate speech and counterspeech during the COVID-19 outbreak. I contributed heavily to our analysis and pre-print as first author [2]. We found that, although hate tends to beget more hate, counterspeech can discourage users from becoming hateful in the first place. *Within a week*, our work was **added to the curriculum at Stanford University** [3], and *within less than a month*, it was **covered by New Scientist** [4].

In my second summer project, I helped build a new and challenging *implicit hate speech* benchmark dataset. With guidance from a political scientist at Georgia Tech, I refined this broad class of hate speech into finer-grained categories like *incitement* and *stereotypical language*. Now I am co-leading our efforts to detect these subtle categories and automatically generate *summary* text to identify the target demographic and explain the message's implied meaning. Moderators could use this tool to better understand why a particularly subtle post was problematic, and then moderators could suggest a more positive revision. With insights on counterspeech in the first project and tools for moderation in the second, my work will facilitate both top-down as well as grassroots efforts to prevent hate speech. Additionally, these projects will guide the research community to move beyond isolated key-word matching and towards more socially and politically grounded understanding of subtle antisocial behaviors.

As a first-year PhD student at Georgia Tech, I have since expanded my set of research questions to complement my previous and ongoing efforts in antisocial computing. I am leading a study on the linguistic framing of police shootings and racial violence in America, which is a highly salient and contentious topic in US political discourse. Previous work has shown that framing, or the highlighting of certain discourse elements, is a linguistic signature of ideological difference. By systematically studying framing, I will be better equipped to understand conflict and predict the downfall of conversations before they turn aggressive. Additionally, I am a collaborator on a separate effort to provide automatic resume evaluation and career advice to socially and economically disadvantaged populations. In this way, I am leveraging NLP techniques not only to counter bad behavior on the internet, but also to promote social good.

Intellectual merit. Throughout my early research career, I have demonstrated commitment to addressing the problem of toxic behavior and promoting civility in online spaces. I believe my history of interdisciplinary work uniquely positions me to effectively leverage the insights of social scientists and other domain experts and to conduct more nuanced and socially responsible research in this space. Furthermore, I know how to work with industry partners to acquire the rich datasets I need to properly situate my understanding of abuse with first-hand reports, user profiles, meta data, and other contextual signals. I have contributed novel insights and methods for measuring the social context of norms, intentions, and interpersonal power dynamics that have been previously overlooked in the field, and most importantly, I have shown that, without these contextual signals, text-based models will be *unable* to capture the nuances of toxic behaviors like cyberbullying [1].

As one of my ICWSM reviewers commented, my "*emphasis on context is important. The recognition that aggressive language is not diagnostic merits assertion until the rest of the world gets it!*" I too believe that my work needs to be amplified, especially in a field that is quickly reaching a saturation point with over 3.3 thousand total papers [5]. Despite extensive effort, cyberbullying and hate speech have *not* yet been solved. My research agenda will move the field in the right direction. As an early career researcher, the GRFP would provide me with the support and visibility I need to make this important and lasting contribution.

Broader Impacts. As the son of two educators, I witnessed from an early age the compounding returns that education can provide. Even before I became a full-time student, I valued the stories of those who had returned to thank my parents for their investment. Now I too am the product of all that my educators and research mentors have invested in me. As a QuestBridge Scholar, my undergraduate education was made possible by the generous support of anonymous donors who believed that low-income and underrepresented students are worth investing in. I am wholeheartedly committed to help provide the same opportunities for others and to help train the next generation of research scientists and innovators.

During my undergraduate years, I enthusiastically dedicated my spare time to teaching, mentorship, and service. As a two-time *Alternative Break* co-lead, I joined the *Volunteer Emory* staff and mentored 10 other co-lead pairs, guiding each one to manage a successful service trip of their own. These trips required months of planning to coordinate with non-profits and to ensure the safety and well-being of the 12 undergraduates who would volunteer their time. My experience in each role taught me critical *budgeting, team building* and *human resource management* skills that have prepared me to manage my own lab, ensuring that each member's needs are met. Additionally, I have years of academic mentoring experience. I have been tutoring my peers since I was in middle school, and in college, I worked closely with 10 and 20 of my peers each semester as a calculus tutor, designing individually-tailored problem sets and offering personal feedback to suit each student's unique learning style. I will continue this commitment to teaching, mentorship, and service, now as a graduate student, and throughout my academic career.

Following the completion of my PhD, I will pursue a university faculty position with close ties with the industry where I believe I will be optimally positioned to collaborate across departmental lines for high-impact research and social good. As a faculty member, I will be willing and eager to mentor students from varying academic backgrounds, not just in Computer Science. Through interdisciplinary collaboration, I believe I can help make the internet a safer, healthier, and more inclusive environment, and I am committed to bringing about this shift now, at a time when cultural polarization is high and misinformation is rampant.

References [1] **Ziems** et al. (2020). *Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification*. In ICWSM. [2] **Ziems**, et al. (2020). *Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis*. arXiv preprint. [3] Jurafsky (2020). *CS 384: Ethical and Social Issues in Natural Language Processing*. Retrieved from *web.stanford.edu/class/cs384*. [4] Lu (2020). *Scams, lies, and online hate*. In New Scientist, Issue 3286. [5] A. Waqas, J. Salminen, S.G. Jung, H. Almerekhi, and B.J. Jansen. *Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate*. In PloS one 2019.